Artificial Intelligence In Medicine 99 (2019) 101690

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Using game theory and decision decomposition to effectively discern and characterise bi-locus diseases

Nassim Versbraegen[a,b,1,*], Aziz Fouché[a,c,1,**], Charlotte Nachtegael[a,b], Sofia Papadimitriou[a,b,e], Andrea Gazzo[a,b,d], Guillaume Smits[a,f,g], Tom Lenaerts[a,b,e,*]

[a] Interuniversity Institute for Bioinformatics in Brussels, ULB-VUB, 1050 Brussels, Belgium
[b] Machine Learning Group, Université Libre de Bruxelles, 1050 Brussels, Belgium
[c] École normale supérieure Paris-Saclay, 94230 Cachan, France
[d] Center for Medical Genetics, Reproduction and Genetics, Reproduction Genetics and Regenerative Medicine, Vrije Universiteit Brussel, UZ Brussel, 1050 Brussels, Belgium
[e] Artificial Intelligence Lab, Vrije Universiteit Brussel, 1050 Brussels, Belgium
[f] Hôpital Universitaire des Enfants Reine Fabiola, Université Libre de Bruxelles, 1020 Brussels, Belgium
[g] Center for Medical Genetics, Hôpital Erasme, Université Libre de Bruxelles, 1070 Brussels, Belgium

ARTICLE INFO

Keywords:
Oligogenic disease
Classification
Game theory
Random forest
Feature interpretation

ABSTRACT

In order to gain insight into *oligogenic* disorders, understanding those involving bi-locus variant combinations appears to be key. In prior work, we showed that features at multiple biological scales can already be used to discriminate among two types, i.e. disorders involving *true digenic* and *modifier* combinations. The current study expands this machine learning work towards *dual molecular diagnosis* cases, providing a classifier able to effectively distinguish between these three types. To reach this goal and gain an in-depth understanding of the decision process, game theory and tree decomposition techniques are applied to random forest predictors to investigate the relevance of feature combinations in the prediction. A machine learning model with high discrimination capabilities was developed, effectively differentiating the three classes in a biologically meaningful manner. Combining prediction interpretation and statistical analysis, we propose a biologically meaningful characterization of each class relying on specific feature strengths. Figuring out how biological characteristics shift samples towards one of three classes provides clinically relevant insight into the underlying biological processes as well as the disease itself.

1. Introduction

With the advent of affordable sequencing technology, a considerable amount of genetic data became available. The scientific community has since employed those data to uncover many aspects of human genetics, one of these being the genetic component of diseases. Most of these efforts so far have been directed at researching single locus (or gene) mutations causing a disease (i.e. monogenic disease). While great advances have been achieved through this approach, many disorders remain unexplained. It is now evident that the cause of some disorders can only be elucidated by expanding the explanatory model to include combinations of variants in different loci. Therefore, efforts are undertaken to develop machine learning approaches that can unravel *oligogenic inheritance*[2] patterns in diseases. Even though the term *locus* refers to a specific region in a chromosome that does not necessarily correspond to a gene, we will use both terms as synonyms, as is often done in the literature.

When the phenotype of a patient is better explained by mutations in two loci/genes rather than by one, it is said to be explained by a *digenic* model [1]. As the term digenic ambiguously refers to either diallelic heterozygous combinations [2] or all potential variant combinations between pairs of variants between two genes [1,3], we will use *bi-locus* to explicitly refer to the latter. Several rare disorders are now known to be caused or modulated in a bi-locus manner; among them *Usher syndrome* [4,5] that causes deaf-blindness, *familial long QT syndrome* [6–8] which eventually provokes fainting or sudden death or *familial hemophagocytic lymphohistiocytosis* [9] known to induce trouble in immune system monitoring. Furthermore, *bi-locus diseases* represent the very

* Corresponding authors at: Université Libre de Bruxelles, Département d'Informatique, Boulevard du Triomphe CP 212, 1050 Brussels, Belgium.
** Corresponding author at: École Normale Supérieure Paris-Saclay, 61 Avenue du Président Wilson, 94230 Cachan, France.
E-mail addresses: nversbra@ulb.ac.be (N. Versbraegen), afouche@ens-paris-saclay.fr (A. Fouché), tlenaert@ulb.ac.be (T. Lenaerts).
[1] These authors contributed equally to this work.
[2] Inheritance involving multiple loci.

https://doi.org/10.1016/j.artmed.2019.06.006
Received 11 September 2018; Received in revised form 21 June 2019; Accepted 30 June 2019

first step towards understanding *oligogenic* and ultimately *complex*[3] inheritance. Complex disorders can arise from several genes and some affect millions of people around the world, such as the well-known *Alzheimer's disease* [10,11] and *multiple sclerosis* [12].

Bi-locus diseases consist of a pool of candidate gene pairs, among which variant combinations are responsible for the disease phenotype [13]. It is important to note that in some cases the candidate pool is broad, such as in *Bardet-Biedl syndrome* for which 43 gene pairs, spread across 12 different loci, have currently been identified [3]. Studying bi-locus diseases in human medicine thus turns out to be tricky; not only are they rare, but finding two patients suffering from the same disorder with a similar genotype (i.e. exactly the same variant combination) is uncommon. DIDA [3] — DIgenic diseases DAtabase, a manually curated database containing classifications and references of bi-locus pathologies found in literature — was developed in order to facilitate the study of such disorders. Within that resource, three types of pairwise or bi-locus combinations associated with diseases have been observed, which we refer to as *Bi-locus Effects* (BE); *True Digenic*, *Modifier* and *Dual molecular diagnosis* instances (see Fig. 1 for their definitions).

Identifying these classes entails associating relevant biological features at different biological scales to the disease type [3]. Indeed, *true digenic* combinations can be handled analogously to recessive diseases (which require a homozygous variant), in the sense that the conjunction of mutations in two genes is needed for the individual to present the phenotype. For *modifier* ones, the major locus determines whether the individual is ill, while the minor one is a modifier of symptom severity or age of onset; therefore, their contributions to a given disease are asymmetric. Finally, *dual molecular diagnosis* combinations, initially present in negligible quantities in DIDA, are essentially responsible for the conjunction of two independent monogenic disorders that occur simultaneously within a single patient [2]. Although the mechanisms are different for each category, it is often nontrivial to know to which class a newly discovered combination belongs, especially when there is no information available about the relatives. Disentangling the spectrum of bi-locus combinations based on a set of descriptive multi-scale features would thus be immensely useful, which is exactly the ambition of the current work. The question thus is which multi-scale features characterize the different BEs and how their interplay can be used to identify each effect type.

The remainder of the paper is organized as follows; we first provide information on the prior work the current study builds on. Employed data sets and methods in the paper are then presented, preceding the results generated by this work. We end with a discussion on these results, aiming first of all to provide biological insights into the three different bi-locus disease types. The work is part of a more extensive pipeline that includes the prediction of the pathogenicity of novel variant combinations.

1.1. Background

A binary machine learning model using random forests [see 14] was previously conceived by Gazzo et al. in order to differentiate between *true digenic* and *modifier* cases. This predictor achieved respectable results, yet as in this first version of DIDA [3] no or very few *dual molecular diagnosis* were present, patterns defining *dual molecular diagnosis* cases could not be extracted [3].

In that work, feature selection was performed to determine the most suitable way to characterize each bi-locus combination available in DIDA. Doing so is nontrivial, as each sample can be represented using numerous features which share complex interactions. Gazzo et al.'s intuition was to use three different biological levels (variant, gene and gene pair level) to describe a gene pair,[4] achieving a relatively good

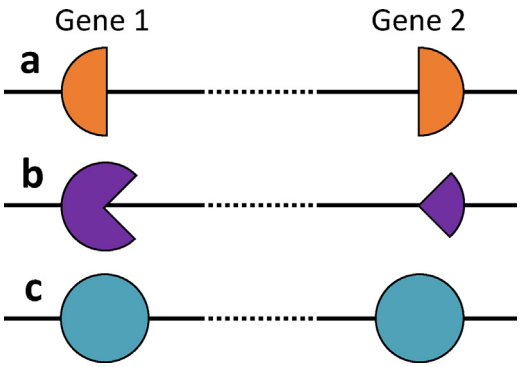

Fig. 1. The three types of bi-locus effects. Combination a, *true digenic* combinations where the simultaneous presence of a pathogenic allele in each gene is necessary for the individual to express the morbid phenotype. **Combination b**, *monogenic plus modifier* combinations where a variant on the *major* gene induces a disease phenotype while a mutation in the *modifier* gene modulates it, either by rendering it more severe or producing an early onset. **Combination c**, *dual molecular diagnosis* combinations where both loci are responsible for either distinct or overlapping phenotypes.

predictive quality [3]. Moreover, inside a bi-locus combination, genes are sorted according to their Gene Damage Index (GDI) [15], which provides quantification of the mutational damage experienced by a gene in the general human population. A gene with a lower GDI is typically better conserved, and thus more probable to be involved in a disease if damaged. Such a gene is invariably put first in the bi-locus combination. This approach significantly improved the predictive quality of the model, which is also confirmed in this work.

Due to the scarcity of *dual molecular diagnosis* combinations in the original version of DIDA, they were grouped together with the *modifier* cases under the *composite* denomination [16]. As a consequence, the earlier predictor solved only the binary classification problem, i.e. separating *true digenic* from *composite* instances. However, recent clinical work by Posey et al. provides a significant expansion of *dual molecular diagnosis* cases, summarizing data on more than one hundred patients that are highly relevant for the current study [2]. 75 of their *dual molecular diagnosis* combinations provided exploitable data, making the *dual molecular diagnosis* class equivalent in size to the *true digenic* and *modifier* classes. Concurrently, DIDA was updated with publications until June 2017 (dida.ibsquare.be), introducing additional *true digenic* and *modifier* combinations. This novel data set triggered the current initiative to produce a predictive model with the capacity to differentiate all three classes in a biologically meaningful manner.

Although predictive quality is in itself important, uncovering the decision process of the classifier is as critical, given that it provides biologically meaningful insights to non-experts in Artificial Intelligence (AI) and Machine Learning (ML) like clinicians or geneticists. Palczewska et al.'s [17] algorithm for *random forest* decomposition is applied here to our predictive model to unravel the differences between the three BEs. Additionally, the explanatory quality of the different features used to construct the model are quantified using *Shapley Index* [18], a method to assess vote importance developed within the context of cooperative game theory. This method considers the prediction process as a payoff-game, whose reward is some measure of the prediction quality [19]. Each feature can be isolated as an agent in order to compute its individual contribution, which is expressed by the Shapley value. Expanding this idea, feature pairs are investigated to locate them on a feature interaction continuum. The pairs associated to low interaction values are called *redundant* (i.e. the pair does not perform better than its individual components). The average ones are *complementary*

[3] Inheritance involving multiple loci plus environment factors.

[4] Recall that a bi-locus combination consists of two genes, and therefore can involve two to four variants.

(*footnote continued*)

(i.e. the performance of one pair compares to the sum of its components), while the high-valued pairs are *synergistic* (i.e. their conjunction yields a better prediction than the sum of the marginal contributions) [20]. Together these two methods will make the model explanatory, ensuring that clinicians and geneticist can use their expertise to assess the BE prediction made by the model.

2. Materials and methods

2.1. Data sets

Two data sets have mainly been used. Whereas the 75 *modifier* and 90 *true digenic* combinations stem from the updated version of DIDA [16], the 75 *dual molecular diagnosis* ones have been retrieved from Posey et al.'s work [2]. The final data set therefore consists of 240 combinations. To avoid the introduction of a batch effect, the only information gathered was *chromosome, position, reference allele and alternative allele* (CPRA) for each variant inside the combinations, as well as the gene names. These values are sufficient in order to subsequently build the data set.

2.2. A three-class classification method using RF

The highly limited size of the data set constrained the number of suitable machine learning methods one can explore. We decided to work with *random forests*, given their desirable performance on small data sets.

Random forest is a relatively recent machine learning ensemble method [21] that relies on the aggregation of multiple decision trees. In essence, it functions by combining decisions provided by an ensemble of such trees, which are one of the most explicit (and hence interpretable) models in machine learning [22]. A decision tree consists of successive data set splits with respect to conditions on features. Each leaf contains a label (or continuous value in the case of regression trees) which gets assigned to each query sample that ends up at that leaf after traversing the tree following a path imposed by its features.

As shown by Gazzo et al., bi-locus combinations involving the same genes tend to share the same bi-locus effect, (although some counter examples exist) [16]. To avoid the introduction of a bias that would result in overoptimistic performance measures, *stratified cross-validation* (also known as *leave one group out*) was performed. We split the data set in subgroups defined by a specific gene pair. Then, we iteratively trained the model on the data set minus one gene pair subgroup, and tested it on the samples taken out. This methodology allowed us to evaluate each sample exactly once and assess the performance of our predictive method.

`RandomForestClassifier` from the *Python3* library `sklearn` was used to carry out the predictions. Each forest consists of 100 trees with a depth going up to 10, using bootstrapping, and using the *Gini* splitting criterion as feature importance measure. Each stratified-cross-validation was performed 100 times (unless specified otherwise) over the 240 gene pairs, depending on the precision relevant to the experiment.

2.3. Selected features

As was demonstrated in [16], a feature set relevant for the task at hand considers three genetic levels: (1) the variant level, where each allele (out of the four total alleles of the combination) is annotated separately; (2) the gene level, which includes insight into the importance of the gene in the cell; (3) the combination level, expressing information about the relationships that exist between the genes or their protein products. Specifically, we use here the following features, visualized also in Fig. 2, to define each of the 240 bi-locus combinations:

- **Variant-level features**, which include a measure of pathogenicity for each allele that was predicted by CADD [23], which empirically showed the best results, and is moreover well accepted by the bioinformatics community and supports many types of mutations.
- **Gene-level features**, consisting of two values for each gene; (1) a boolean value indicating whether it is considered essential in mice [as determined in 24]; (2) the probability for the gene to be recessive, and thus not dominated by a pathogenic allele [25].
- **Combination-level features**, which is a higher-level boolean value indicating whether gene products are involved in the same pathway according to KEGG [26] or REACTOME [27] databases.

Different zygosity states are implicitly captured in the feature vector representation; for instance, when CADD values are only provided for the first allele of gene A and the first allele of gene B (and thus both second alleles are assigned -1), the combination is heterozygous.

The impact of additional features was investigated, such as allelic state (how many mutant alleles are involved in a combination), co-expressed genes or biological distance [28] instead of pathway, but did not show an improvement in prediction results. It may be because these features are uncorrelated with bi-locus effects, or because the information they carry is redundant with respect to already integrated features. Therefore, we decided not to take them into account in the proposed model. We also tried to use different pathogenicity predictors, namely DEOGEN [29] and DANN [30], but they yielded lower results in terms of prediction quality.

Out of the 240 digenic combinations, 3 CADD values were missing, as well as 23 essentiality A or B values, and 78 recessivity A or B values. Unknown values for a feature were imputed by using the mean of the corresponding explanatory variable – no significant difference was observed between using the mean or the median. Wild-type variants were assigned a pathogenicity score of -1, corresponding to the minimal CADD score value.

2.4. Feature analysis using game theory

Given the limited number of features, it is possible to evaluate all possible feature combinations with a moderate computational cost. Such an evaluation allows one to assess the importance of a feature in the decision process, both individually and in combination with the other features. To achieve this goal some principles from cooperative game theory can be used, as mentioned in Section 1: we define a payoff-game using the features (represented by $\mathcal{F}$) as a set of agents and the prediction quality in terms of sensitivity and specificity as the rewards [19]. As there are three classes to predict and there are two measurements for predictive quality, i.e. sensitivity and specificity, each evaluation implies six different games and thus six different rewards. To make the process more understandable, these six measurements are combined into one unique value. Details are provided below.

First of all, every feature coalition (i.e. set of features to consider) $S \subset \mathcal{F}$ is assessed by carrying out 50 stratified-cross-validations using the developed predictor, leading to the knowledge of the prediction capabilities for all coalitions S. As there are 9 features that can either be considered or discarded, there are $2^9 - 1$ possible coalitions S, as the game cannot be played with the empty coalition. In the following, results will be addressed as $\mathrm{sen}_{\mathrm{class}}^{\mathrm{coalition}}$ and $\mathrm{spe}_{\mathrm{class}}^{\mathrm{coalition}}$ to designate the sensitivity and specificity respectively, obtained by a given coalition for a given class. Class names will be abbreviated TD for *true digenic*, MM for *modifier* and DMD for *dual molecular diagnosis*. The empty coalition is considered to have no inferential power, yielding

$$\mathrm{sen}_{\mathrm{class}}^{\varnothing} = 0 \quad \text{and} \quad \mathrm{spe}_{\mathrm{class}}^{\varnothing} = 0.$$

Then, Shapley [18] was used to analyze feature importance by investigating the contribution of each given feature $f_i \in \mathcal{F}$ to the overall prediction quality. First, it is necessary to define a mapping

$$v\colon S \subset \mathcal{F} \longrightarrow \mathbb{R},$$

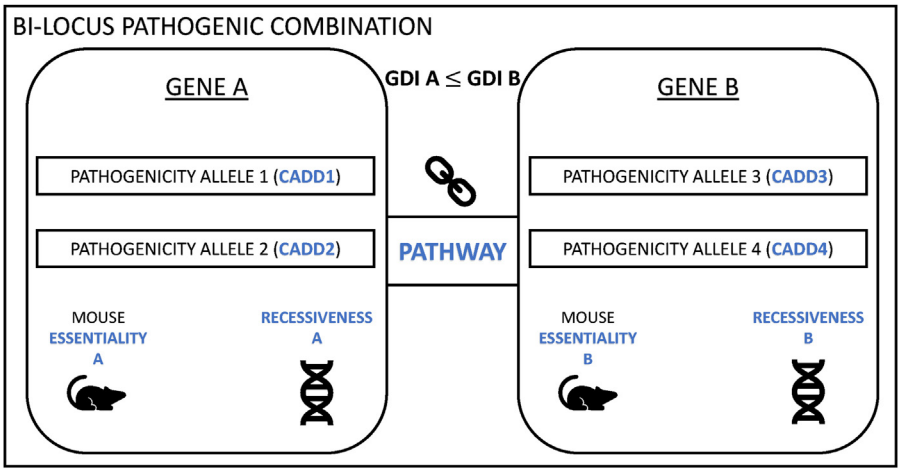

Fig. 2. Diagram representing the architecture of a bi-locus combination, with the three levels. Features, namely CADD1, CADD2, CADD3, CADD4, Essentiality A, Essentiality B, Recessiveness A, Recessiveness B and Pathway are emphasized in blue. GDI allows for ordering the two genes inside the combination. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

that associates each coalition to a reward, based on the six metrics considered (sensitivity and specificity for each BE). Let $C = \{TD, MM, DMD\}$ be the set of the three classes. In our case, the coalition S is the feature subset, and the reward $v(S)$ is the geometric mean of the 3 sensitivities and the 3 specificities, defined as

$$v(S) = \left(\prod_{c \in C} \text{sen}_c^S \times \text{spe}_c^S\right)^{1/6} \tag{1}$$

Shapley value $\phi_{\mathcal{F}}$ provides a relative value indicating how much an agent contributes to a coalitional reward. Adapting it to the feature selection process gives the formula for a given feature $f_i \in \mathcal{F}$

$$\phi_{\mathcal{F}}(f_i) = \sum_{S \subset \mathcal{F} \setminus f_i} \frac{|S|!(|\mathcal{F}| - |S| - 1)!}{|\mathcal{F}|!}(v(S \cup f_i) - v(S)) \tag{2}$$

The `SHAP` package [31] was used to compute individual feature contributions. The package was chosen for its efficiency through heuristics, thus avoiding exhaustive result computations over $2^{\mathcal{F}}$.

2.5. Interaction analysis using game theory

Along with the individual feature contributions, it is also interesting to examine the interaction between features. To that end, a technique similar to Kaufman et al.'s [20] was applied, aiming to identify how features act together, i.e. to examine whether they act synergistically or whether they exert no influence on each other. The principle is analogous to the Shapley value, with an extended expression. Here, the underlying intuition is to reward a feature pair if they both contribute equally and produce good results, and to penalize a feature pair if it behaves asymmetrically (i.e. one agent consistently contributing more than the other), or if the pair tends not to improve coalitions it takes part in (*dummy* pair).

For two distinct features f_i and f_j, their *joint contribution* $\gamma_{i,j}$ represents how good the coalitions they both belong to, perform. It is defined using (Eq. (1)) as follows:

$$\gamma_{i,j} = \sum_{S \subset \mathcal{F} \setminus \{f_i, f_j\}} \frac{|S|!(|\mathcal{F}| - |S| - 2)!}{(|\mathcal{F}| - 1)!}(v(S \cup \{f_i, f_j\}) - v(S)) \tag{3}$$

On the other hand, *marginal contributions* $\delta_{i,j}^i$ where f_i acts without f_j and $\delta_{i,j}^j$ where f_j acts without f_i are defined using (2) as

$$\begin{aligned} \delta_{i,j}^i &= \phi_{\mathcal{F} \setminus f_j}(f_i) \\ \delta_{i,j}^j &= \phi_{\mathcal{F} \setminus f_i}(f_j), \end{aligned} \tag{4}$$

and aim to catch the asymmetry between the two features of interest.

Joint (Eq. (3)) and marginal (Eq. (4)) contributions are combined to assess the interaction $I_{i,j}$ between f_i and f_j.

$$I_{i,j} = \gamma_{i,j} - \delta_{i,j}^i - \delta_{i,j}^j \tag{5}$$

As the `SHAP` package does not support interaction analysis with `scikit-learn` random forests yet, algorithms to compute these values were implemented in *Python3*. Corresponding source code can be found in this project's GitHub repository.[5]

2.6. Classifier interpretation

Palczewska et al. discussed an algorithmic method specific to random forests aiming to get, for a given prediction, quantitative information for each feature about its contribution to the decision process [17]. As their approach is used in the current work to analyze the feature relevance, we will briefly describe the algorithm.

The algorithm, which can be generalized to k classes, is run on each trained tree $T_l = (V, E)$ within the forest independently. Results are eventually averaged to get an insight of the full forest decision process.

Let

$$\text{val}: V \to [0, 1]^k$$

be a mapping associating each tree node to the proportions of samples belonging to each class situated on the leaves below. The *branch gradient* $g_{v,w}$ is then defined for each descending branch $(v, w) \in E$ as

$$g_{v,w} = \text{val}(w) - \text{val}(v). \tag{6}$$

A general bias b, proper to the trained tree, is defined as $b = val(r)$, where r is the root of T_l.

A sample $x \in \mathcal{X}$ traversing the tree T_l following a path $\{v_0, ..., v_n\}$, can be seen as unfolding all the $g_{v_i,v_{i+1}}$ (see Eq. (6)) along that path.

$$T_l(x) = \text{val}(v_n) = b + \sum_{i=0}^{n-1} g_{v_i,v_{i+1}}. \tag{7}$$

Where, when considering the full traversal, one gets

$$\sum_{i=0}^{n-1} g_{v_i,v_{i+1}} = \text{val}(v_n) - \text{val}(v_0) = \text{val}(v_n) - b.$$

As $val(v_n)$ is exactly the predicted vector given the input x, this approach basically decomposes the decision process in gradients $g_{v,w}$ plus a bias inherent to the data set. Each gradient can be linked to the feature $f_i \in \mathcal{F}$ responsible for the splitting of v. Therefore, a prediction $T_l(x)$ where x drops down the path $\{v_0, ..., v_n\}$ can be dissected as a sum of feature contributions

[5] https://github.com/oligogenic/DIDA_SSL.

$$T_l(x) = b + \sum_{i=1}^{|\mathcal{F}|} \text{contrib}_l(f_i, x), \tag{8}$$

where $contrib_l(f_i, x) = \sum_{j=0}^{n-1} \mathbb{1}_{(f_i \text{ splits } v_j)} g_{v_j, v_{j+1}}$ is the feature contribution of f_i for the tree T_l in the x prediction. It is easy to verify this definition of $T_l(x)$ to be equal to $val(v_n)$ using Eq. (7), as exactly one feature splits one node, and each node is split by one feature.

Finally, the global contribution of the feature f_i to the forest decision process can be obtained by averaging $contrib_l(f_i)$ over all trees T_l. This algorithm was implemented in *Python* and can be found in the *GitHub*[5] associated with the current study.

3. Results and discussion

3.1. A three-class classification model to disentangle bi-locus types

Classic cross-validation methods cannot be properly used on small data sets as they deprive the predictor of large slices of information it already lacks. Furthermore, making the same gene pair appear both in the training and the testing set introduces a bias, as shown by [16], which must be avoided to gauge predictor's performance on new combinations. Therefore, a stratified-cross-validation based on gene pairs has been performed to evaluate models in an unbiased manner, described in detail in Section 2.

To evaluate prediction quality, predictor performances are compared to those produced by a randomized model picking one class among the three with probability 1/3, simulated 5000 times. As displayed in Fig. 3, *dual molecular diagnosis* cases appear to be the easiest to identify, both in terms of sensitivity and specificity. This high accuracy makes sense from a biological perspective: indeed, these results reveal that the 3-class predictive model is able to separate instances more closely linked to monogenic diseases from those requiring bi-locus genotypes. 80% of sensitivity and specificity for *dual molecular diagnosis* combinations reveals that the used features are clear markers for discriminating effectively between two monogenic and a bi-locus genetic model. The rest of the bi-locus effects (*true digenic* and *modifier*) also lead to good prediction levels, performing significantly better than a random classifier. Differences in performance with prior work are due to the addition of the third class (*dual molecular diagnosis*), increasing the number of potential class mismatches and, therefore, decreasing results overall. These results show that Gazzo et al.'s [16] features space remains robust, even with the addition of a third bi-locus effect. However, prediction quality seems to have reached its limit in this framework, suggesting an increase in samples, or the exploration of other features may be worthwhile to differentiate consistently between the *true digenic* and *modifier* classes.

3.2. Characterization of bi-locus types based on variant strengths

Random forest decisions can be comprehensively dissected by a systematic approach, as discussed in Section 2. The prediction is decomposed as a sum over all features of individual contributions that can be either positive (the feature increases the class probability) or negative (the feature decreases the class probability). This technique reveals that, even if all features globally influence the decision process, the *dual molecular diagnosis* class is mainly recognized via the CADD1 and CADD3 features (Fig. 4, bottom panel), corresponding to the respective pathogenicities of one mutated allele in gene A, and one mutated allele in gene B — *dual molecular diagnosis* combinations are mostly di-allelic. This observation was further confirmed, as a predictor running only on these two features is able to achieve very good *dual molecular diagnosis* classification accuracy, with a sensitivity of 0.74 ± 0.01 and a specificity of 0.73 ± 0.01. These high results clearly indicate that CADD1 and CADD3 features are highly informative in order to characterize *dual molecular diagnosis* combinations. We must nonetheless note that they are not sufficient to reach both sensitivity and specificity obtained when all features are taken into account, indicating other features also contribute to the decision process. It is also interesting to note that this limited feature space is not sufficient to identify all bi-locus effects. Indeed, when considering only di-allelic pathogenicity, both *modifier* and *true digenic* classes present a specificity and sensitivity under 0.52, which can be explained by their strong similarity in terms of their pathogenic mutation profile.

Expanding these observations to a statistical analysis shows that *dual molecular diagnosis* combinations present higher pathogenicity with respect to the di-allelic genetic profile, consisting of CADD1 and CADD3, with unfortunately a high standard deviation (Table 1). Nonetheless, a Student's *t*-test suggests the difference between *dual molecular diagnosis* di-allelic profiles and other BEs to be significant, especially for CADD1 as shown in Table 2. Therefore, we conclude that *dual molecular diagnosis* combinations have a significantly more pathogenic di-allelic profile than other bi-locus effect combinations,

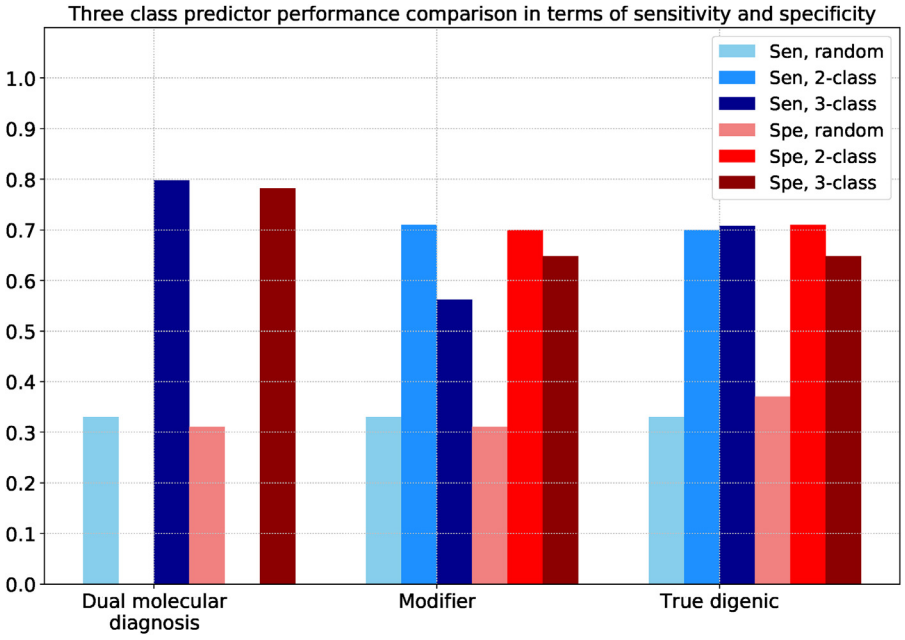

Fig. 3. Prediction quality in terms of sensitivity and specificity for the three bi-locus effect classes. Comparison of a naive randomized 3-class predictor that randomly chooses a class among the three, regardless of the input, the 2-class predictor of the previous study [16] that does not support *dual molecular diagnosis*, and the 3-classes predictor developed in our work. The naive 3-class predictor is only used as a baseline, to compare the predictor we propose. Sensitivity and specificity are used as performance metrics. Differences in results in terms of *true digenic* and *modifier* classification compared to prior work are explained by the addition of the third *dual molecular diagnosis* class, increasing the number of possible confusions therefore decreasing results in absolute. Note also that some *modifier* combinations are suspected to actually be *dual molecular diagnosis*.

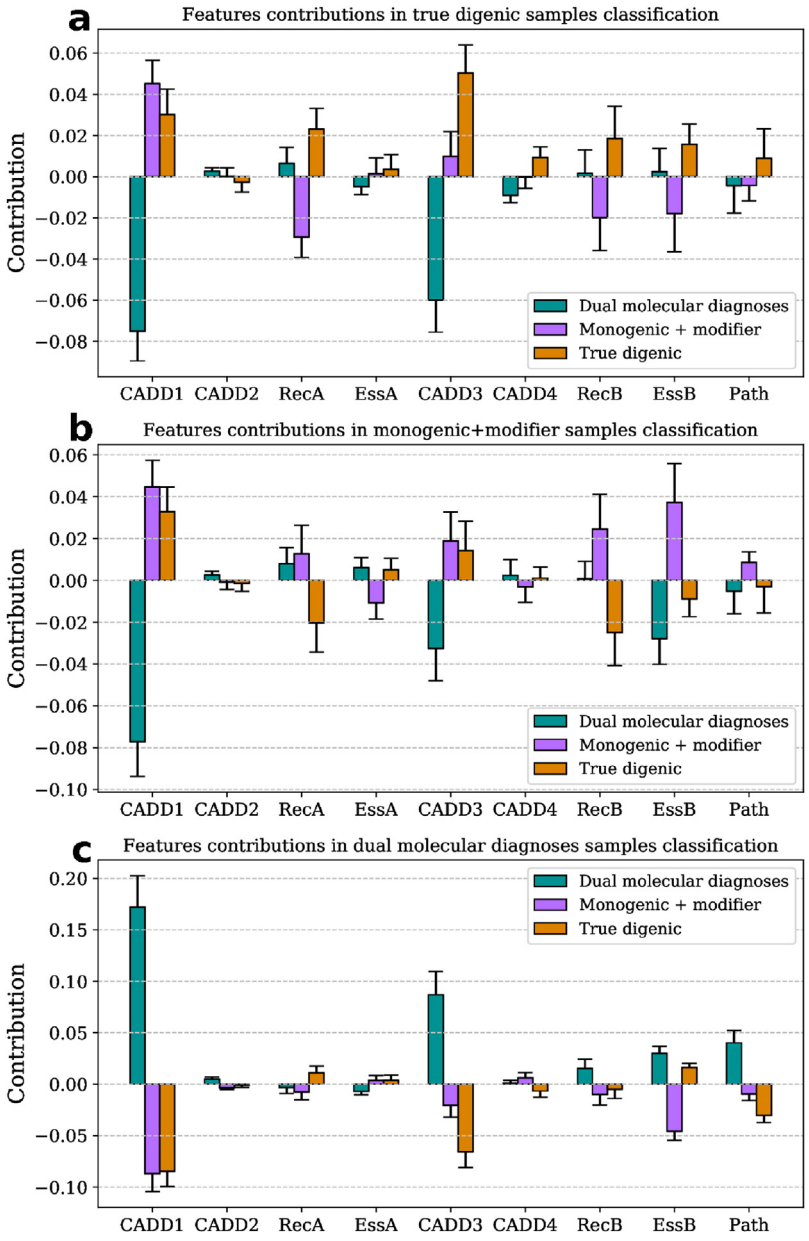

Fig. 4. Each figure represents the contribution of each feature in the decision process, when classifying (a) *true digenic* samples, (b) *modifier* samples, (c) *dual molecular diagnosis* samples. A positive contribution of a feature for a class influences the decision positively towards the class (we refer to this as *pulling*). Conversely, a negative contribution of a feature for a class orientates the decision against the class (*pushing*). Bars represent the mean contribution while error bars represent the standard error, where each trial corresponds to a training of a random forest on the whole dataset minus a sample, then a prediction on this sample using the trained model.

Table 1
Di-allelic profile of each bi-locus effect. We can see *dual molecular diagnosis* to be highly pathogenic compared to other bi-locus effects.

Bi-locus effect	CADD1 (mean, std)	CADD3 (mean, std)
True digenic	3.34 ± 2.13	3.79 ± 2.96
Modifier	3.36 ± 1.66	3.42 ± 1.65
Dual molecular diagnosis	**6.22** ± 2.73	**5.50** ± 3.72

Table 2
As pathogenicity differences were denoted between *dual molecular diagnosis* combinations and other BEs, *p-values* were computed to determine the likelihood for this singularity to be due to chance. **Top triangular:** *p-value* with respect to CADD1 feature. **Bottom triangular:** *p-value* with respect to CADD3 feature. The CADD1 difference between *dual molecular diagnosis* combinations and the rest is the most significant feature difference.

Class1	Class2		
	True digenic	Modifier	Dual diagnosis
True digenic	1	0.94	2.97×10^{-12}
Modifier	0.34	1	1.86×10^{-12}
Dual diagnosis	1.26×10^{-3}	2.01×10^{-5}	1

characterizing genetic profiles with stronger pathogenic capabilities.

On the other hand, Fig. 4 shows an asymmetry between *true digenic* and *modifier* in terms of *dual molecular diagnosis* class probability contribution. While in *true digenic* combinations, CADD1 and CADD3 values equally repel samples from the *dual molecular diagnosis* class, in *modifier* there is a clear asymmetry between these two features, with CADD3 being less polarized towards the *true digenic* class, and contributing more to the *dual molecular diagnosis* class than the CADD1 feature. As *dual molecular diagnosis* combinations are known to have strong pathogenic variants able to trigger a disease phenotype. This feature contribution asymmetry may be correlated with the strength of the major variant in the case of *modifier* combinations. We should note that considering CADD values simply depicts how damaging the mutation is for the protein, without taking into account its negative impact at the scale of the whole organism. Therefore, strong variations in proteins

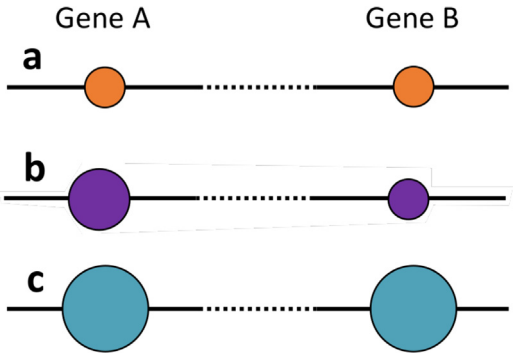

Fig. 5. Schematic representations of the three bi-locus effects in terms of deleterious capabilities. **Combination a**, *true digenic* with two weak (i.e. low pathogenicity) variants. **Combination b**, *monogenic plus modifier* with one mild variant and one weak variant. **Combination c**, *dual molecular diagnosis* with two strong variants.

that do not play an important role may be less deleterious than small variations on essential ones. Using higher-level features such as gene essentiality and recessiveness, the model seems to be able to catch the deleteriousness of the major gene. We discuss these interpretations further in Section 4.

These observations bring a first characterization of bi-locus diseases in terms of variant pathogenicity (Fig. 5). Statistical analyses as well as phenotypic insights show that *dual molecular diagnosis* tend to contain stronger variants. On the other hand, the random forest interpretation suggests that in *true digenic* cases, both variants are equally important, while there is a clear imbalance in the importance of the variants involved in *modifier* cases, as CADD3 orients the predictor towards the *dual molecular diagnosis* class. Even if no statistical argument can be made to support this last observation, the coherence between tree interpretation and phenotypic observations suggests that the full feature framework (i.e. including features from the three different levels) better represents how deleterious the variant combination will be for the organism in comparison to using exclusively information at the variant level, thus more effectively distinguishing the *major* locus from the *modifier* one.

Fig. 7 shows how *dual molecular diagnosis* combinations are indeed shifted towards the east (and north to a lesser extent) of the plot in the space of all instances used in this study, containing more damaging variants than those present in the two other bi-locus classes.

3.3. Second-order analysis shows pathway feature's high synergy

Two features are said to be *synergistic* for a prediction if they share a better than complementary relationship, meaning that each characteristic carries information whose relevance is highly improved when combined with the other (their combination being more than the sum of the parts). To investigate these interactions, a game theoretical approach, more extensively discussed in Section 2, was adopted. Using this quantification, each feature is tested both individually and against all the others, generating a 2D array whose diagonal represents the individual contributions, i.e. Shapley values. Although the assessed feature importance is globally similar to the results one would obtain with the Gini Importance score [22], slight differences can be observed, which are most likely due to differences in how the scores are calculated. The Shapley approach is preferred given that it was shown to be consistent [32]. The intersection of a row and a column outside of the diagonal shows the synergy value between the corresponding features. These relative values indicate how two features perform together with regard to the global prediction quality, and can be observed in Fig. 6.

This plot reveals that the *pathway* feature, although presenting average relevance on its own, turns out to be critical when coupled with other features. Indeed, by considering the *pathway* feature in

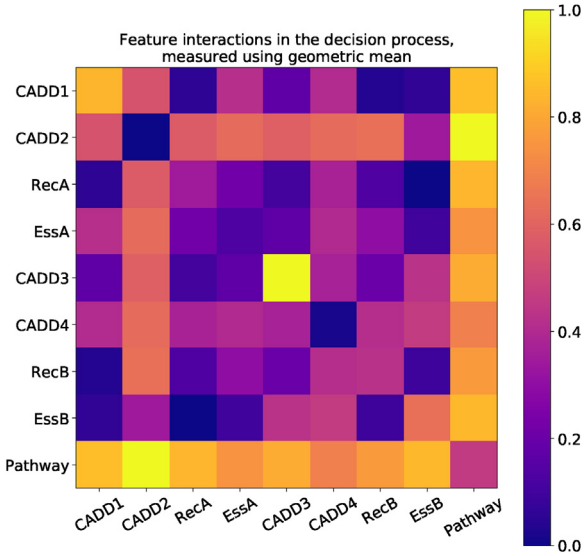

Fig. 6. Bi-dimensional figure presenting both features individual contributions to the decision process (*diagonal*), and pairwise contributions. Prediction quality p is assessed using the geometric mean of sensitivity and specificity with respect to each class. The figure shows on one hand the individual importance of CADD1, CADD3, and EssB, and reveals on the other hand the high synergy of *pathway* with all features. Values have been linearly interpolated between 0 and 1 for clarity purposes, and are therefore relative.

characterizing the di-allelic profile, *true digenic* and *modifier* classifications improve by 12% and 25% respectively, whereas *dual molecular diagnosis* classification accuracy increases by a modest 2%. Although this feature analysis approach is limited to cases with few features due to the computational cost in $O(n^2)$ with n the amount of features, it turns out to be very powerful as it allows an analysis including second-order feature contributions (i.e. taking into account feature interactions, and not only individual contributions). Furthermore, this cost can be mitigated using dimension reduction techniques, while the original features can be traced back from the new basis. This is especially relevant, as it reveals that *true digenic* and *modifier* combinations are more easily differentiated when *pathway* information is present, and when this information is combined with other features. This notion of common pathway needs to be considered carefully, as genes implicated in bi-locus inheritance must somehow take part in a common biological process, either at the level of transcription regulation, or in (indirect) protein-protein interaction for instance. This is attested by the fact that different combinations of the two genes may vary the phenotype.

A statistical analysis shows that on average, *modifier* combinations tend to be more pathway-related (57%) than *true digenic* ones (29%), but with very high standard deviations (close to 50%). Therefore, solely this feature is not informative enough for differentiating between these two classes. *dual molecular diagnosis* combinations rarely appear to be *pathway*-related (a mean of 11%), thus appearing as a less significant signature than the information at the variant level. This observation reinforces the notion of loci independence in *dual molecular diagnosis* combinations, as two different monogenic disorders co-occur in the same individual. However this insight should be treated with caution as when a bi-locus combination has a *pathway* value of 0, this does not necessarily indicate that there is no common pathway between the two genes, but may mean that a common pathway is not known yet.

This pathway importance can also be distinguished in Fig. 4. The pathway feature is active in *true digenic* and *dual molecular diagnosis* classification, capturing the fact that genes inside *dual molecular diagnosis* combinations are rarely biologically linked together.

We can also observe in the plot of Fig. 6 that CADD2, as well as CADD4 (though in a lesser manner), have a high synergy with many of the used features. CADD2 and CADD4 represent the CADD scores of the

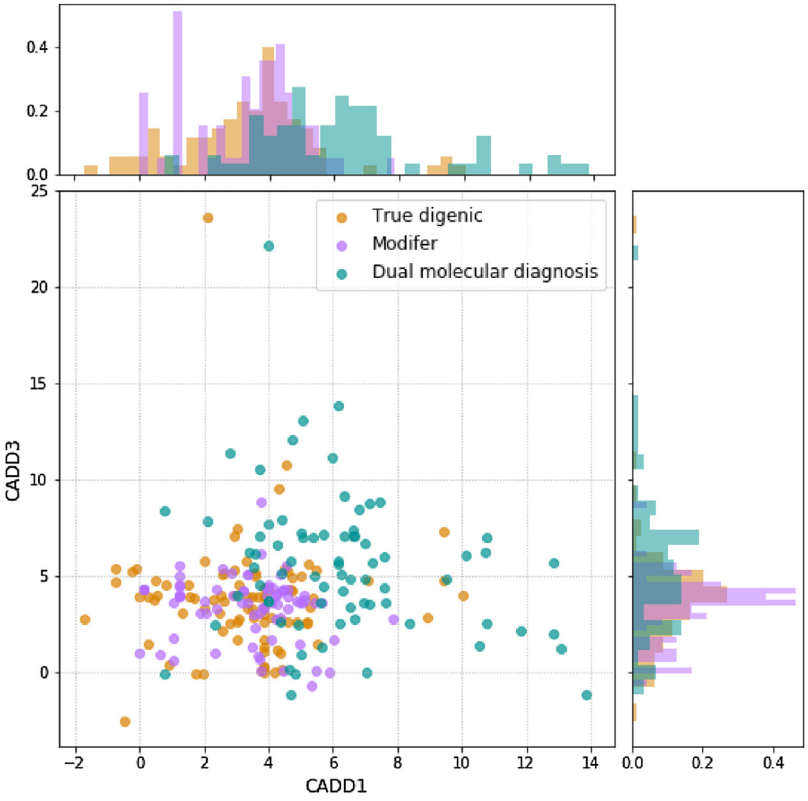

Fig. 7. Combinations spreading in the space with respect to the pathogenicity of their first and third allele, showing relative high di-allelic pathogenicity of *dual molecular diagnosis* combinations. Histogram plots show the density of combinations of each class along CADD1 (horizontal) and CADD3 (vertical) axes.

second variant allele available for each gene. This means that they implicitly provide zygosity information, carrying values for homozygous variants, and in general for the more complex (tri- and tetra-allelic) variant combinations. Their usefulness in detecting the digenic effect of a combination seems to also depend on the information present in the rest of the features, and especially on the recessiveness and essentiality of the involved genes, as well as the pathway membership, for which they show a certain level of synergy. This observation intuitively makes sense, as the effect of a homozygous or heterozygous variant can also depend on the inheritance and expression pattern of the involved gene, and thus these features seem to mutually contribute to the decision process. As CADD4 corresponds to the second allele of the variant in the gene with a less severe deleteriousness, it could be expected, and is actually shown, that the synergy is less profound compared to the case of CADD2, which corresponds to the variant lying in the gene with a more severe deleteriousness.

3.4. Bi-locus effect landscape of disease instances — BEspace

To further gain insight in how features influence the relation between instances of different classes, we developed an online visualizer that provides an interactive plot of all the instances used to train the predictor. The plot uses t-SNE [33,34] that enables dimensional reduction, projecting the multi-dimensional space onto a two dimensional space. Instances of each class are labeled by a distinct color. Users can select the features (i.e. taking them into account for the projection in two dimensional space or not), as well as the different classes (by clicking on them in the legend). Furthermore, summary information about the features and classes is presented when hovering the selector circle over an area. The tool, thus, provides an intuitive, visual way to assess aggregated feature sets. Though not as precise as predictor validations, one can evaluate the influence of features by looking at how instances of classes are grouped together: a more homogeneous clustering indicates a better separability, while a heterogeneous cluster indicates that the currently enabled features might not be ideal to discriminate between the different classes. BEspace is available at http://bespace.ibsquare.be.

4. Conclusions

By analyzing the prediction process at different levels, *dual molecular diagnosis* combinations have been shown to display a strong biological signature, allowing the machine learning model to recognize them effectively. Since *modifier* and *true digenic* combinations share a similar bi-locus profile, it is more difficult to accurately discriminate between them. Notwithstanding this limitation, the predictor is able to identify these two classes well, based on both sensitivity and specificity metrics, indicating that the feature space contains some relevant pieces of information. Given the limited size of the data set, adding more features may lead to overfitting. Therefore, entirely reconsidering which features to use, redefining the bi-locus profiles from scratch or adding new samples might lead to prediction improvement between those two bi-locus classes. Nevertheless, discarding variant-level features may decrease the *dual molecular diagnosis* recognition quality.

As the interpretation of results is key in bioinformatics, decision decomposition methods are, in general, necessary. The game theoretical approach used in this work presents both advantages and disadvantages, as it scales quite well in terms of data size by using batches to simulate the whole data set, but does not scale in terms of feature quantity. Nonetheless, this approach may be applied even in high dimensions, after the application of dimension reduction techniques, such as PCA and t-SNE. Therefore, this approach remains relevant, allowing an easier interpretation of the model. The combination of this approach with the tree interpretation, and the assessment of the coherence between them, reinforces the results' validity and explains why the machine learning model is so effective in recognizing *dual molecular*

diagnosis combinations; indeed, the tree interpretation algorithm works at the decision-level, while the synergy analysis does so at the result-level.

We propose a meaningful bi-locus spectrum characterization by combining different conceptual layers of information, a statistical analysis and a random forest interpretation (Fig. 5). This characterization associates the ability of a variant to trigger a disease phenotype with a pathogenic strength that can be assessed either at the *variant-level* with the use of predictors like CADD [23], or in a more complex way by considering features at different levels. A solid framework is defined with this characterization and can be extended in future work to improve the classification task. For *modifier* combinations, it is interesting to note that the gene with the lower GDI tends to be the *major* gene, as lower GDI indicates more conservation, and thus lower tolerance to mutations.

The exploration of synergistic features revealed important and meaningful information. The game theoretical approach highlights the importance of the *pathway* feature in discriminating different bi-locus effects. While *dual molecular diagnosis* combinations were easily differentiated on the variant level due to the strong pathogenicity of the involved variants, *true digenic* and *modifier* combinations showed the necessity of higher levels of information in order for them to be classified correctly; *pathway* information was among those features, indicating interactions between the genes at the protein level and therefore subsequent consequences of the mutations on these interactions. This shows that *bi-locus diseases* work on an interactive level and thus features representing this interaction were most significant when classifying *true digenic* and *modifier* combinations. An explanation for this could be that *true digenic* combinations need both variants to disrupt pathways in order for the phenotype to appear, whereas *modifier* combinations already show a mild phenotype with only the major gene affected by the mutation, meaning that the second alteration in the minor gene could affect the same pathway and causes worse disease symptoms or earlier age of onset. Possible *true digenic* combination mechanisms have been proposed, such as the need for several proteins in a protein complex to be mutated in order for that complex to be destabilized, or the presence of numerous disruptions occurring at different steps of a pathway [35]. Additional information about the level of interaction between the genes or the gene products could fine-tune the classification of *bi-locus diseases*.

Both prediction and statistical analysis showed that *dual molecular diagnosis* combinations are different from the other two bi-locus effects. Their high di-allelic pathogenicity clearly indicates their monogenic capabilities, allowing them to trigger symptoms by themselves without requiring another locus. Furthermore, their tendency not to belong to a common pathway reinforces the idea they can also be present independently. As such, one may wonder how relevant it is to consider them under the bi-locus denominator. Confounding things further, the symptoms observed for *dual molecular diagnosis* combinations may be split between *distinct* ones, where each disease affects a distinct tissue of the organism and *overlapping* ones, where diseases are located in the same organ or tissue and share similar phenotypes, requiring a more careful interpretation of the predicted results.

More specifically, the overlapping *dual molecular diagnosis* combinations require cautious analysis. Since they affect the same tissues of the organism, they may interact in a common pathway. Therefore, a biologically relevant relationship between the two genes could exist. We can even hypothesize that this relationship reinforces the severity of the symptoms, by a snowball effect. Therefore, if some overlapping combinations could implicate two completely unrelated genes, others may include both the symmetry of *true digenic* combinations, and the modifier effect of *modifier* combinations.

Delving into the features effect on the models allowed us to derive qualitative characteristics and insights into the different classes, which in turn yielded comprehensible biological information, leading to increased understanding of the molecular biology underlying the different classes.

The machine learning model that was conceived during this study can now be used either as part of a pipeline for bioinformaticians or as an investigative tool for physicians. Considering its ability to effectively identify different types of bi-locus combinations, its usage can be essential to attain the right diagnosis in clinic when new bi-locus genotypes are identified. From a machine learning perspective, a bi-locus combination severity predictor may have usage of such a model, conditioning its decision on the predicted input combination's bi-locus effect.

Funding

This work was supported by the ARC project Deciphering Oligo- and Polygenic Genetic Architecture in Brain Developmental Disorders; the European Regional Development Fund (ERDF) and the Brussels-Capital Region-Innoviris within the framework of the Operational Programme 2014–2020 through the ERDF-2020 project ICITY-RDI.BRU [27.002.53.01.4524]; a Fonds de la Recherche Scientifique (F.R.S) – FNRS Fund for Research Training in Industry and Agriculture (FRIA); a Vrije Universiteit Brussel, Reproduction and Genetics and Regenerative Medicine (RGRG) Cluster, Reproduction and Genetics Research Group.

Conflict of interest statement

Authors declare no conflict of interest.

Supplementary materials

Data sets and scripts are available at the address https://github.com/oligogenic/DIDA_SSL.

Acknowledgements

Great thanks go to the IB^2 team interested in oligogenic diseases, for their useful comments and crucial suggestions. Thanks to the École Normale Supérieure Paris-Saclay for giving A. Fouché the opportunity to come to Brussels.

References

[1] Defrise-Gussenhoven E. Hypothèses de dimérie et de non-pénétrance. Hum Hered 1962;12(1):65–96.
[2] Posey JE, Harel T, Liu P, Rosenfeld JA, James RA, Coban Akdemir ZH, et al. Resolution of disease phenotypes resulting from multilocus genomic variation. New Engl J Med 2017;376(1):21–31.
[3] Gazzo AM, Daneels D, Cilia E, Bonduelle M, Abramowicz M, Van Dooren S, et al. Dida: a curated and annotated digenic diseases database. Nucleic Acids Res 2015;44(D1):D900–7.
[4] Ebermann I, Phillips JB, Liebau MC, Koenekoop RK, Schermer B, Lopez I, et al. Pdzd7 is a modifier of retinal disease and a contributor to digenic usher syndrome. J Clin Investig 2010;120(6):1812–23.
[5] Bonnet C, Grati M, Marlin S, Levilliers J, Hardelin J, Parodi M, et al. Complete exon sequencing of all known usher syndrome genes greatly improves molecular diagnosis. Orphanet J Rare Dis 2011;6(1):21.
[6] Westenskow P, Splawski I, Timothy KW, Keating MT, Sanguinetti MC. Compound mutations: a common cause of severe long-qt syndrome. Circulation 2004;109(15):1834–41.
[7] Paulussen A, Matthijs G, Gewillig M, Verhasselt P, Cohen N, Aerssens J. Mutation analysis in congenital long qt syndrome – a case with missense mutations in kcnq1 and scn5a. Genet Test 2003;7(1):57–61.
[8] Millat G, Chevalier P, Restier-Miron L, Da Costa A, Bouvagnet P, Kugener B, et al. Spectrum of pathogenic mutations and associated polymorphisms in a cohort of 44 unrelated patients with long qt syndrome. Clin Genet 2006;70(3):214–27.
[9] Zhang K, Chandrakasan S, Chapman H, Valencia CA, Husami A, Kissell D, et al. Synergistic defects of different molecules in the cytotoxic pathway lead to clinical familial hemophagocytic lymphohistiocytosis. Blood 2014. blood-2014.
[10] Yoshiiwa A, Kamino K, Yamamoto H, Kobayashi T, Imagawa M, Nonomura Y, et al. α1-antichymotrypsin as a risk modifier for late-onset alzheimer's disease in Japanese apolipoprotein eε4 allele carriers. Ann Neurol 1997;42(1):115–7.
[11] Kamboh MI, Sanghera DK, Ferrell RE, DeKosky ST. A4poe* 4-associated alzheimer's disease risk is modified by α1-antichymotrypsin polymorphism. Nat Genet 1995;10(4):486.

[12] IMSGC, et al. Evidence for polygenic susceptibility to multiple sclerosis – the shape of things to come. Am J Hum Genet 2010;86(4):621–5.
[13] Schäffer A. Digenic inheritance in medical genetics. J Med Genet 2013;50(10):641–52. https://doi.org/10.1136/jmedgenet-2013-101713.
[14] Breiman L. Random forests. Mach Learn 2001;45(1):5–32.
[15] Itan Y, Shang L, Boisson B, Patin E, Bolze A, Moncada-Vélez M, et al. The human gene damage index as a gene-level approach to prioritizing exome variants. Proc Natl Acad Sci USA 2015;112(44):13615–20.
[16] Gazzo A, Raimondi D, Daneels D, Moreau Y, Smits G, Van Dooren S, et al. Understanding mutational effects in digenic diseases. Nucleic Acids Res 2017;45(15):e140.
[17] Palczewska A, Palczewski J, Robinson RM, Neagu D. Interpreting random forest models using a feature contribution method. 2013 IEEE 14th international conference on Information reuse and integration (IRI). 2013. p. 112–9.
[18] Shapley LS. A value for n-person games. Contrib Theory Games 1953;2(28):307–17.
[19] Cohen S, Dror G, Ruppin E. Feature selection via coalitional game theory. Neural Comput 2007;19(7):1939–61.
[20] Kaufman A, Kupiec M, Ruppin E. Multi-knockout genetic network analysis: the rad6 example. Computational systems bioinformatics conference, 2004. CSB 2004. Proceedings. 2004 IEEE. 2004. p. 332–40.
[21] Ho TK. Random decision forests. Proceedings of the third international conference on document analysis and recognition, 1995, vol. 1. 1995. p. 278–82.
[22] Breiman L. Classification and regression trees. Routledge; 1984.
[23] Kircher M, Witten DM, Jain P, O'roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 2014;46(3):310.
[24] Georgi B, Voight BF, Bućan M. From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. PLoS Genet 2013;9(5):e1003484.
[25] MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science 2012;335(6070):823–8.
[26] Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res 2013;42(D1):D199–205.
[27] Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, et al. The REACTOME pathway knowledgebase. Nucleic Acids Res 2013;42(D1):D472–7.
[28] Itan Y, Mazel M, Mazel B, Abhyankar A, Nitschke P, Quintana-Murci L, et al. Hgcs: an online tool for prioritizing disease-causing gene variants by biological distance. BMC Genomics 2014;15(1):256.
[29] Raimondi D, Gazzo AM, Rooman M, Lenaerts T, Vranken WF. Multilevel biological characterization of exomic variants at the protein level significantly improves the identification of their deleterious effects. Bioinformatics 2016;32(12):1797–804.
[30] Quang D, Chen Y, Xie X. Dann: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics 2014;31(5):761–3.
[31] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017. p. 4768–77.
[32] Lundberg SM, Erion GG, Lee SI. Consistent individualized feature attribution for tree ensembles. 2018. arXiv preprint arXiv:1802.03888.
[33] van der Maaten L. Accelerating t-sne using tree-based algorithms. J Mach Learn Res 2014;15(1):3221–45.
[34] van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9(November):2579–605.
[35] Badano J, Katsanis N. Beyond Mendel: an evolving view of human genetic disease transmission. Nat Rev Genet 2002;3:779–89.